Environmental Indicators through Artificial Neural Networks

Environmental Indicators through Artificial Neural Networks

Jesús Silva[1], Alexa Senior Naveda[2], Hugo Hernández Palma[3], William Niebles Núñez[4], Luis Jiménez Rodríguez[5]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2] Universidad de la Costa, Barranquilla, Atlántico, Colombia
[3] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[4]Universidad del Sucre, Sincelejo, Sucre, Colombia.
[5]Universidad Colegio mayor de Cundinamarca, Cundinamarca, Colombia.

Email: jesussilvaUPC@gmail.com

Abstract. Indicators are the most important management tool for environmental monitoring. Environmental indicators condense the information and simplify the approach to environmental phenomena, which are often complex, and makes them very useful for communication. The usefulness of these indicators consists of providing relevant information, summarized in the form of concise and illustrative statements for decision making, both for the organization's management and for the rest of the members. The prediction of limit values, together with the potentialities offered by the recommendation system based on ontology make this system a powerful tool for supporting decision-making in the Environmental Management process with a wide possibility of generalization in the business sector.

1.Introduction

The usefulness of these environmental indicators consists of providing relevant information, summarized in the form of concise and illustrative statements for decision making, both for the organization's management and for the rest of the members. Therefore, they ensure a rapid assessment of the main improvements and weaknesses in the company's environmental protection, for those who make decisions [1]. For this reason, it is necessary to use environmental indicators for measuring the behavior of the organization in this area, facilitating communication and condensing environmental information. The use of indicators, in turn, contributes to an improvement in the management of environmental knowledge.

In order to manage knowledge, its representation is decisive, which represents the process of structuring knowledge about a problem in a way that is easier to solve. In order to promote management and specifically the representation of knowledge, Semantic Technologies (ST) are increasingly used. Within ST, ontologies are currently one of the most widely used Forms of Knowledge Representation Systems (KRS) [2].

In this sense, [3] developed an "ontology-based system for knowledge management with environmental indicators" (SIGCIA) with the objective of managing the environmental knowledge that is inferred from the historical storage of the environmental business indicators.

For this reason, the study proposes the use of Artificial Neural Networks (ANN) to predict the limit value of the indicator from its historical storage. Among the potentialities that ANNs offer is that they do not need a human expert for extracting knowledge.

2. System Based on Ontology of Environmental Indicators

Today, one of the most commonly used forms of knowledge representation are ontologies, which offer dissimilar advantages for the modeling, generation, distribution and use of knowledge produced and accumulated in organizations [4]. Given these advantages for knowledge management, ontologies are widely used to manage the large volumes of environmental information that result from this process, mainly from the historical storage of environmental indicators.

The SIGCIA system is based on the *OntoEnvironmental* ontology, in which the environmental indicators governing the corporate environmental management process are modeled [5]. The software based on this system of indicators allows their calculation considering that the indicator must have predefined its limit value (which is manually defined by the Environmental Management Specialist in correspondence to the tacit agreement on the historical behavior of the indicator). In response to this action, the system compares the value and limit value of the indicator. If the value of the indicator is greater than the limit value, the system declares that the indicator is altered and, through the inference machine, recommends possible causes, possible environmental impacts and mitigation actions [6].

However, there is yet deficiency for the potential improvements that the implementation of the SIGCIA system offers for the correct performance of the environmental management process in the organizations. It is difficult for the Environmental Management Specialist to establish limit values due to the fact that indicators reflect different areas of the entity (e.g., energy area, transport), therefore, the value is established in a subjective way. As a result, a bad decision to establish a limit value restricts the potential offered by the system by not making recommendations in a timely way [7] [8].

3. Knowledge Discovery in Databases

Large volumes of data and information that are currently handled have resulted in the need to develop techniques and tools to assist man to extract useful information, knowledge and patterns of stored data. Knowledge Discovery in Databases (KDD) comes to meet this need.

According to [9], KDD is defined as "The non-trivial process of identification, in the data, of valid, novel, understandable and potentially useful patterns". KDD is a computing area that attempts to exploit the enormous amount of information by discovering representative and useful patterns, extracting knowledge that can assist a human to perform tasks in a more efficient and satisfactory way.

The phases that this process goes through are shown in general terms by [10] [11] [12]:

- Selection: develops an understanding of the problem domain and the data that will be used in the knowledge discovery task.
- Pre-processing and transformation: they include all activities for the construction of the final dataset. These tasks include selection of records, attributes, data cleaning, treatment of missing values, among others. Data transformation is also performed in the format required by the selected data mining tool. This task consumes between 35% and 20% of the time.
- Data mining (MD): is the determination of the discovery task to be performed (classification, regression, grouping, ...) and the application of one or more algorithms of that task, in order to discover hidden patterns in the data. This task occupies between 15% and 20% of the project's completion time.
- Interpretation and evaluation: the patterns discovered are interpreted and evaluated, so it is sometimes necessary to return to the previous steps, which implies repeating the process, perhaps with other data, algorithms, goals and strategies. This step can be aided by visualizations and contributes to eliminating redundant or irrelevant patterns.

3.1 Data selection

Every KDD project has its origins in the request of a client who wants to improve some of his processes using the historical data. In order to take full advantage of these data, it is necessary that the implementers of this type of project know and understand the data. The historical storage of each environmental indicator constitutes a dataset. This indicator needs to have its limit value calculated in order to know when it is altered and recommend possible causes, possible environmental impacts and

mitigation actions. The KDD scheme was applied to perform the calculation, with the aim of finding a model that would allow this value to be obtained as accurately as possible [13] [14] [15].

In order to carry out this research, the data referring to the monthly energy consumption indicator of a Construction Company (CC) are available. This information contains a history of approximately 5 years (from December 1, 2014 to July 1, 2019). The CC manages 6 parameters to record the monthly electricity consumption in its files, as shown in Table 1.

Table 1. Description of the dataset

Attribute	Value
Date	data
global active power (kilowatt)	actual
global reactive power (kilowatt)	actual
voltage (volt)	actual
global intensity (ampere)	actual
consumption (watt/hora)	actual

Figure 1 shows the behavior of the instances of dataset. It can be noted that, in most months, the energy consumption is between 1700- 2555 (watt/hour).

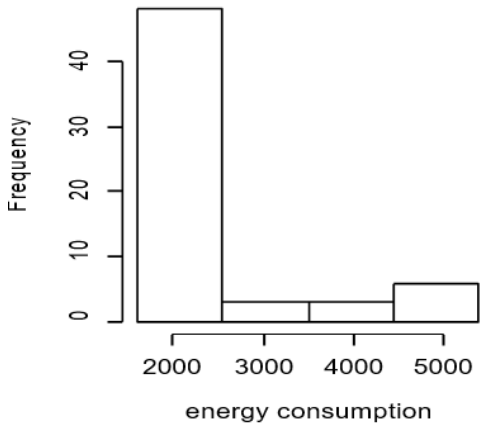

Figure 1. Distribution by instance (consumption attribute)

3.2 Attribute Selection

Attribute Selection (AS) can be defined as the process of obtaining the most representative *n* attributes of the original *N* from the elimination of redundant and irrelevant *N*s. In a more formal way, the objective is to select a subset of *S* attributes of the original space *A* with class *C*, such that $P(C \mid S) \approx P(C \mid A)$, that is, to obtain better or equal predictive performance through the elimination of noisy and redundant attributes [16].

There are several criteria for grouping the AS algorithms. One of them is the application mode, according to which they can be seen as filter or wrapper. In this study, the way for applying the algorithms for selecting attributes was by means of the wrapper criterion. Enveloping strategies are those that use the classifier precision to evaluate the subsets of the space. This strategy offers better results, since in a previous step to the classification, the Learning Algorithm chooses the attributes that best represent the knowledge for its construction. However, it is highly expensive.

Five algorithms of enveloping strategies were used: Linear Regression (LR), Multi-layer Perceptron (MLP), M5P, K-nearest neighbors (K-nn) and M5Rules (M5R) [17].

3.2.1 Linear Regression
Regressive analysis is a technique used for inter and extra polar observations, which can be classified as linear or non-linear regression. A regression model is when the response variable and the explanatory variables are all quantitative. It allows determining the mathematical model or equation that best represents the existing relationship between the analyzed variables [18].

3.2.2 Multi-Layer Perceptron
An Artificial Neural Network (ANN) is a computational model that aims to simulate the functioning of the brain. The learning process of an ANN with Multi-layer Perceptron topology consists of determining the weights that allow the underlying knowledge to be coded in the data [19]. This consists of varying the weights according to some learning rule until they are constant, so it is said that the network has learned. Its good predictive performance is given by the high noise tolerance of the data and the ability to capture complex relationships between attributes and the class.

3.2.3 M5P
In the case of the M5P algorithm, it is an issue of obtaining a model tree (a linear model that predicts the value of the class), although it can be used to obtain a regression tree since this is a specific case of a model tree [20].

3.2.4 K-Nearest Neighbors
It is a simple algorithm that stores all available cases and classifies new cases on the basis of a similarity measure (distance functions). K-nn has been used in statistical pattern recognition, estimation and, already in the early 1970s, as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the most common class among its K-nearest neighbors, measured by a function of distance. If K = 1, then the case is simply assigned to the class of its nearest neighbor [21].

3.2.5 M5Rules
M5Rules performs a span regression, with each span determined from a regression tree. It implements base routines to generate M5 Models of trees and rules. The original M5 algorithm [11].

3.3 Data Mining
In this research, an experimental study was prepared to explore the behavior of ANNs in datasets where their class type is continuous. The Weka tool [3] was used for the execution of experiments, which is a software developed at the University of Waikato (New Zealand) under GNU (General Public License), and is characterized by its architectural independence.

An experimentation scheme based on cross validation is used to ensure greater statistical robustness. This proposal consists of a cross validation procedure with ten partitions with a run as proposed by [21]. It is used as an evaluation parameter:

- Correlation coefficient
- Absolute mean error

Correlation is the statistical technique that studies the problem of measuring the intensity or degree of relationship that exists between the variables being researched. The Correlation Coefficient is a value between -1 and 1 that indicates the linear relationship that exists between two variables. The absolute mean error measures the average magnitude of errors in a set of forecasts, regardless of their direction. It measures accuracy for continuous variables. An experiment was performed using the Wrapper strategy selection algorithms mentioned above and as regression algorithms: LR, MLP, M5P, K-nn and M5R. The results are shown in Tables 2 and 3.

3.4 Interpretation and evaluation

In Table 2 and Table 3 the highlighted values are the algorithms with best correlation coefficient and mean absolute error respectively. It can be noted that the MLP regression algorithm with the Wrapper MLP strategy attribute selector presents the highest correlation coefficient in Table 2 and the lowest absolute mean error in Table 3.

Table 2. Result of correlation coefficient

	Wrapper					
LR		**MLP**	**M5P**	**K-nn**	**M5R**	**-**
LR	0,685	0,671	0,566	0,1785	0,696	0,611
MLP	0,785	0,931	0,799	0,235	0,868	0,821
M5P	0,898	0,873	0,915	0,796	0,898	0,836
K-nn	0,799	0,886	0,724	0,687	0,632	0,854
M5R	0,811	0,790	0,885	0,799	0,821	0,965

Table 3. Absolute mean error result

	Wrapper					
LR		**MLP**	**M5P**	**K-nn**	**M5R**	**-**
LR	757	798	741	833	747	752
MLP	465	336	425	786	325	336
M5P	398	375	336	588	436	375
K-nn	335	298	398	347	395	396
M5R	474	368	447	325	347	387

Figure 2 shows the CC electricity consumption over five years. In this trend graph, consumption is presented by months. The blue color represents the actual consumption, while the red color is the consumption predicted by the MLP algorithm. It shows that the error of the classifier is low.

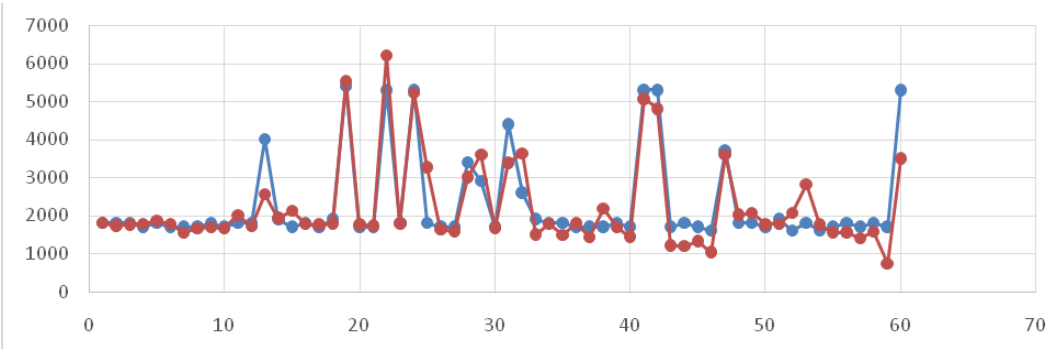

Figure 2. Energy Consumption Forecast. Red refers to the forecast and blue to the actual value.

4. Conclusions

In the study of the attribute selection algorithms with the Wrapper strategy to select the methods to be applied in the study, it was possible to verify that redundant and irrelevant attributes existed, due to the fact that the one with the best results was Wrapper (MLP), eliminating the "Intensity" attribute. The study of five regression models from different branches showed that the Multi-Layer Perceptron regression algorithm showed the best performance in terms of measured parameters, correlation coefficient and mean absolute error.

The integration of the Multi-Layer Perceptron regression algorithm into the SIGCIA system allows the prediction of the limit value of the energy consumption indicator that was the data set selected in the

current investigation. This facilitates the work of the Environmental Management Specialist because the system makes recommendations in a timely manner and favors decision making in this regard. The results obtained with the application of the Multi-Layer Perceptron regression algorithm to the set of data taken from the Construction Company (CC) regarding the energy consumption indicator show that, for this indicator in other organizations, the aforementioned algorithm can be generalized.

References

[1] Cios, K. J., & Kurgan, L. A. (2000). Trends in Data Mining and Knowledge Discovery. (Dm), 1-26.

[2] Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[3] Demsar, J. (2006). Comparison of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research, vol. 7: 31*.

[4] Hutt, S.; Gardener, M.; Kamentz, D.; Duckworth, A.; D'Mello, S.: Prospectively Predicting 4-year College Graduation from Student Applications. Proceedings of the 8th International Conference on Learning Analytics and Knowledge, pp. 280-289 (2018)

[5] Ahuja, R.; Kankane, Y.: Predicting the probability of student's degree completion by using different data mining techniques. Fourth International Conference on Image Information Processing (ICIIP), pp. 1-4 (2017)

[6] Martins, L.; Carvalho, R.; Victorino, C.; Holanda, M.: Early Prediction of College Attrition Using Data Mining. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1075-1078 (2017)

[7] IHOBE. (1999). *Guía de Indicadores Medioambientales para la Empresa*. Berlin: Ministerio Federal para el Medio Ambiente, la Conservación de la Naturaleza y la Seguridad Nuclear.

[8] Russell, S.; Norvig, P.: Artificial Intelligence A Modern Approach. Pearson Education 3rd Ed, pp. 705 (2010)

[9] Makhabel, B.: Learning Data Mining with R. Packt Publishing 1st Ed, pp. 143 (2015)

[10] Witten, I.; Frank, E.; Hall, M.; Pal, C.: Data Mining Practical Machine Learning Tools and Techniques. Elsevier 4th Ed, pp. 167-169 (2016).

[11] Bucci, N., Luna, M., Viloria, A., García, J. H., Parody, A., Varela, N., & López, L. A. B. (2018, June). Factor analysis of the psychosocial risk assessment instrument. In International Conference on Data Mining and Big Data (pp. 149-158). Springer, Cham.

[12] Gaitán-Angulo, M., Viloria, A., & Abril, J. E. S. (2018, June). Hierarchical Ascending Classification: An Application to Contraband Apprehensions in Colombia (2015–2016). In Data Mining and Big Data: Third International Conference, DMBD 2018, Shanghai, China, June 17–22, 2018, Proceedings (Vol. 10943, p. 168). Springer.

[13] Viloria, A., & Lezama, O. B. P. (2019). An intelligent approach for the design and development of a personalized system of knowledge representation. Procedia Computer Science , 151 , 1225-1230.

[14] Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[15] Viloria, A., Bucci, N., Luna, M., Lis-Gutiérrez, J. P., Parody, A., Bent, D. E. S., & López, L. A. B. (2018, June). Determination of dimensionality of the psychosocial risk assessment of internal, individual, double presence and external factors in work environments. In International Conference on Data Mining and Big Data (pp. 304-313). Springer, Cham.

[16] Bishop, C. (1995). Extremely well-written, up-to-date. Requires a good mathematical background, but rewards careful reading, putting neural networks firmly into a statistical context. *Neural Networks for Pattern Recognition*.
[17] Pretnar, A. The Mystery of Test & Score. Ljubljana: University of Ljubljana. Retrieved from: https://orange.biolab.si/blog/2019/1/28/the-mystery-of-test-and-score/ (2019).
[18] Castellanos Domíngez, M. I., Quevedo Castro, C. M., Vega Ramírez, A., Grangel González, I., & Moreno Rodríguez, R. (2016). *Sistema basado en ontología para el apoyo a la toma de decisiones en el proceso de gestión ambiental empresarial*. Paper presented at the II International Workshop of Semantic Web, La Habana, Cuba. http://ceur-ws.org/Vol-1797/
[19] Yasser, A. M., Clawson, K., & Bowerman, C.: Saving cultural heritage with digital make-believe: machine learning and digital techniques to the rescue. In Proceedings of the 31st British Computer Society Human Computer Interaction Conference (p. 97). BCS Learning & Development Ltd. (2017).
[20] Castellanos Domínguez, M. I., & Grangel González, I. (2013). *Las ontologías, su uso para la gestión del conocimiento medioambiental.* Paper presented at the III Taller Internacional la Matemática, la Informática y la Física en el Siglo XXI, Holguín.
[21] Khelifi, F. J., J. (2011). K-NN Regression to Improve Statistical Feature Extraction for Texture Retrieval. *IEEE Transactions on Image Processing*, 20, 293-298.